Data Mining Homework Report

In this homework, I have implemented DBSCAN based on chapter 10 pseudocode algorithm in lecture book. I have implemented all my algorithms with python and created a Jupyter notebook. I have used insurance csv dataset file for this study which is I have downloaded from kaggle siteInitially I have explored dataset columns rangeses and scaled them to affect at the same degree to output. Then I selected charges column as a first feature, fort he second feature I observed charges feature's relations between other features by visualizing. According to scatter graphs best feature is age for clustering and for the consistent and accurate output, I have filtered dataset based on female gender and taked only those rows. Now, I will select epsilon and minimum points value for my function. I have used elbow method to select best epsilon value and plot it. In elbow method I have used 2-knn algorithm to find distance between samples. After selecting epsilon value, I have selected minimum point value by trying diffirent number with epsilon value and visualized them. First, for crosscheck I have give the values and dataset to python library DBSCAN function. Then, I have called my function for these values and compared with each other, they were same. Lastly, I have tried my function 3 times with diffirent minimum points number and visualized them. Depending on the changes; If two graphs epsilon values are same, greater minimum points value has more outliers. If two graphs minimum points values are same, greater epsilon value has more cluster labels.

Zeliha Erim

151044065